User-intuitive Explanations for Increasing the Transparency of Autonomous Agents

Francesca Stramandinoli[1*], Brett Israelsen[1], Peggy Wu[1], Kishore K. Reddy[1], Franklin R. Tanner[2], Laura Strater[2]

[1]Raytheon Technologies Research Center

[2]Raytheon Intelligence & Space

*** Correspondence:**
Corresponding Author
francesca.stramandinoli@rtx.com

Keywords: Explainable and Trusted AI, Black-box Explanation, Interpretability Methods, Model-agnostic Explainability, Interpretable Features, Concept-based Explanation, Neuro-symbolic AI.

Abstract

In high tempo, high stakes military applications of Artificial Intelligence (AI) and Machine Learning (ML), operators need to rapidly understand the strengths and limitations of their AI/ML aides, so that the combined human + AI team can provide decision superiority. As AI agents become more sophisticated, operators need transparency in these advanced systems much like the ability to build mental models of their human collaborators to ultimately judge their appropriate use.

A key challenge in making AI explainable to humans who are not computer or data scientists, and who do not have the time or tools to understand the inner workings of the AI system, is to map hidden layers of machine reasoning to semantics that are human interpretable to gain more insight into AI recommendations so that operators may either have more confidence in AI results or know when to override it. To address this challenge of aligning semantics of the machine generated explanation with human interpretation, we first describe the creation of a surrogate white-box approach as a stand-in, and subsequently describe the concept of a semantic alignment method. We describe a use case of Automatic Target Recognition (ATR) and illustrate why current explainer approaches are inadequate in achieving machine transparency and discuss research needs.

1 Introduction

The NORAD and U.S. Northern Command (USNORTHCOM) collect and coordinate a worldwide system of sensors to monitor for aerospace and maritime threats to the United States and Canada. The network of satellites, ground-based radar, airborne radars, and other assets provide a tremendous amount of data. Making sense of this data requires advanced capabilities such as Automatic Target Recognition (ATR) powered by AI/ML. However, autonomous systems alone do not provide the value of human experience and intuition. Further, our adversaries have been monitoring our methods to deter, compete, and conduct war over the past decades. To ensure information dominance and decision superiority, there is a need to combine the ingenuity and agility of human expertise with the pattern recognition and scalable processing power of AI/ML. Human-Autonomy Teaming (HAT) presents unique challenges. Just like human teams, operators need the ability to continuously assess and reassess another team member's performance within different contexts, whether the other team member is human or digital. Current AI/ML methods lack the inherent mechanisms that human teams have in order to represent their levels of competence and trustworthiness. Worse, they lack the ability to be transparent to a human operator. We adopt the National Academies of Sciences (NAS)'s definition of transparency [1] as "*the understandability and predictability of the system*" (Endsley, Bolte, and Jones, 2003, p. 146 [2]), including the AI system's "*abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process*" (Chen et al., 2014a, p. 2 [3]).

Intelligent systems based on opaque AI/ML methods might hide potential issues inherited by biased data or lead to adopting decisions that we do not fully understand, or even worse, that violate ethical principles. Indeed, very often for achieving high performance in prediction, recommendation, and decision-making support, the adoption of complex models that hide the logic of their internal processes is required. Such models are often referred to as black box. A large amount of work that aimed at enabling end users to better understand, trust, and effectively manage artificially intelligent systems has been conducted under the Defense Advanced Research Projects Agency (DARPA) eXplainable Artificial Intelligence (XAI) program [4]. The XAI program envisioned three general approaches for improving explainability while maintaining a high level of learning performance: i) deep explanation approaches to learn features or representations that can make the inner workings of a deep learning model more transparent; ii) interpretable model approaches to develop models that are inherently more explainable and more introspectable for machine learning experts; and iii) model induction approaches that experiment with any given ML model - such as a black box - to infer an approximate explainable model.

This work reviews existing methods for explaining black box models with a focus on techniques that can generate user-intuitive explanations. We describe the work done towards a prototype implementation for generating explanations for AI/ML models that operate on a single modality of data and discuss why current explainer approaches are inadequate in achieving machine transparency and discuss research needs. The use case driving this investigation is Automatic Target Recognition (ATR) where a human analyst is working with the aid of an AI/ML target recognition algorithm.

This paper is outlined as follows. The rest of this section is dedicated to background and foundational concepts, and it describes the particular use case we have had in mind while performing this work. Section 2 is focused on an overview of interpretability methods in use. Section 3 describes the current prototype implementation for explanation generation and research challenges that need to be addressed in pursuit of creating detailed, effective, explanations of black box machine learning models. Section

4 outlines metrics that have been proposed for explainability. Finally, Section 5 contains conclusions, discussion, and possible future research directions.

1.1 Background and Foundational Concepts

In this section we introduce the definitions of transparency, explanations, and interpretability adopted in this work. As noted above, we adopt the National Academies of Sciences (NAS)'s definition of transparency [1] as "*the understandability and predictability of the system*" (Endsley, Bolte, and Jones, 2003, p. 146 [2]), including the AI system's "*abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process*" (Chen et al., 2014a, p. 2 [3]). In other words, a model is transparent if a human can maintain an accurate mental model of how the AI/ML system works. One way to improve model transparency is to provide valid and comprehensible explanations to the human. Following the definition in [5] "*an explanation is an "interface" between humans and a decision maker that is at the same time both an accurate proxy of the decision maker and comprehensible to humans*". In order to create explanations, interpretability methods should be leveraged. We adopt Doshi-Velez and Kim's definition of interpretability as "*the ability to explain or to present in understandable terms to a human*" and extend upon it to add "*without any additional machine processing*" [6]. In other words, an interpretable model is inherently self-explanatory by an operator knowledgeable of the subject. We view interpretability as a continuum that examines the extent to which a process or model can be understood by a human without any assistance (if a user viewed the model alone could they make sense of it? This is discussed further in [37]). An example of an interpretable system is a rule-based system where heuristics used by the model are documented and made available to the human interpreter. This understanding of models is critical for alignment with DoD's Ethical Principles of Traceable and Reliable Artificial Intelligence [7]. Interpretability may not be required in all systems, especially those that can complete their tasks without human intervention [6] (e.g., from toaster to GPS systems). However, creating interpretable systems is necessary in high stakes safety critical human-in-the-loop systems where the ability to understand the machine's reasoning can identify misalignments in objectives and in turn improve a user's ability to make appropriate trust-based decisions.

The design of an interpretable system requires taking several factors into account. An important aspect to consider is the user who is interested in receiving the explanations and why he/she needs an explanation. Identifying the user group (e.g., decision maker, person affected by the decision, designer of the system) helps determining the level of details and type of information to include in the explanation. Another important aspect to consider are the questions that an interpretable system should answer. Different types of AI / ML systems can answer different types of questions (e.g., why, why not, what, what if, how, how confident, etc.,). For example, an ML system trained to solve a classification problem can answer why the model predicts a specific outcome, while a model trained to solve a decision-making task can explain what caused a specific event, showing causal relationships. Moreover, an explanation needs to be contextualized according to when the user needs it (e.g., during onboarding, regular interactions, system errors / malfunctioning, etc.,). In our current work we focus on explainable systems for decision makers that need to have answers to why questions during regular interactions with the system.

Interpretability methods can be classified along several dimensions [8]. An important distinction can be made based on the type of algorithm it can be applied to. Model-specific approaches are specifically designed with a model type in mind and therefore they generally require detailed information about the model; perhaps even require access to the individual parameter weights or other similar details. In contrast, model-agnostic approaches are designed to operate on any model. They only consider the

inputs to a model and corresponding outputs. Another important dimension along which to classify interpretability methods is the scale of interpretation. Global explanations give a high-level overview of how the model performs over its entire mathematical domain. Generally speaking, global explanations sacrifice detailed local precision in favor of global perspective. Local explanations instead are generated for specific examples. They are meant only to be applicable to single predictions of a model and not to offer insights for how the model might classify other examples. The type or modality of data that interpretability methods can be used with, is another crucial factor to consider. *Modality* refers to a particular way in which information is encoded to be presented to a perception/learning system. The following are some examples of different modalities of data: RGB image, infrared image, audio, video, text corpus, GPS trajectory, and time-series sensor data. Typically, models operate on a single modality of data, but they can also be designed to incorporate many modalities at the same time. Lastly, methods for interpretability can be classified based on whether interpretability is achieved by restricting the complexity of the model (intrinsic) by creating a "white-box" or by applying methods that analyze the model after training (post-hoc).

In the ATR use case that will be discussed in Section 1.2, local, post-hoc explanation of specific instances is needed to help operators verify that the identified target is, in fact, the correct one. Simplistic versions of ATR utilize only a single modality of data (typically RGB images) but being able to utilize more modalities of information is critical to improving ATR accuracy. In such a system it is useful to be able to use a model-agnostic explanation algorithm since state-of-the-art algorithms will be in a constant state of change and could even be proprietary. A model-agnostic approach to explanations would ensure the explanation system could remain separate and facilitate implementation in such an architecture and help reduce overall maintenance.

1.2 Use Case for Technology Development

This research is driven by the Automatic Target Recognition use case that involves combining the capability of an ATR system with the expertise of human operators; the main goal is to improve the overall target recognition performance by using the strengths of both the human operator and AI/ML algorithms.

A typical ATR use case can include a variety of sensors from different domains (space, air, and ground sensors) collecting multi-modal data. Sensor data, for example, can include RGB or Electro Optical imagery. Because of the shear amount of incoming data, sensor feeds are processed by AI algorithms to highlight potential targets. AI algorithms could potentially come from different vendors with varying degrees of accuracy and reliability. Human analysts (e.g., signal and /or geospatial analysts) then analyze the targeting data generated by the AI systems assessing whether the evidence supports or contradicts their hypothesis and make the target nominations.

While performing this work, we consider a scenario in which the ATR system must process imagery data to automatically detect the presence of aircrafts. To this end, we leverage an opensource dataset, called Rareplanes, consisting of 253 Maxar WorldView-3 satellite scenes spanning 112 locations and 2,142 km^2 with 14,700 hand-annotated aircraft data [9]. Figure 1 is an example of some images from the Rareplanes dataset with aircrafts identified by red boxes (i.e., bounding boxes). The dataset can be used to create models that predict the number of engines an aircraft has, the type of wings, the number of tail fins, or the role (civil/military transport, fighter, et cetera), among other things. A pre-trained ML model for detecting aircraft role is also included with the dataset [9]. The Rareplanes dataset and utilities are useful for investigating black box explanations.

Figure 1 -- Example images from the Rareplanes dataset. The top 2 rows show examples of real satellite images, the bottom 2 rows are synthetic images [9].

2 Interpretability Methods to Explain Black Box Models

There are many survey papers on the broad topic of explainable AI [10],[11], and the more specific topic of explanation of black box models [5],[12],[13],[14],[15]. This section provides an overview of selected interpretability methods available in the literature that are model-agnostic and post-hoc. Here "model-agnostic" indicates that we want to be able to apply these methods to any generic model that is encountered (i.e., black box model). And "post-hoc" indicates that we are seeking an explanation of a prediction that has been made by the black box model. Our current focus is on interpretability methods acting on imagery data as required by the use case under current consideration (Section 1.2). However, some of the techniques we describe in the following sections might be suitable for other data modalities. In Sections 2.1 and 2.2 we provide an overview of methods that yield to different types of explanations including among others saliency maps, that are the most widespread visual explanation methods, and concepts attribution.

2.1 Methods based on Saliency Map

Before describing some of the interpretability techniques that can generate post-hoc explanations for black box models based on imagery data, we provide a definition of saliency map. Given an image that we want to explain, a saliency map is an image in which a pixel's brightness represents how salient/important the pixel is [15]. LIME (Local Interpretable Model-agnostic Explanations) [16] and SHAP (Shapley Additive Explanation) [17] are model-agnostic approaches that provide explanations of a model prediction by ranking the importance of individual features to a specific prediction, although, as it will be explained below, they accomplish this in slightly different ways. For imagery

data, feature importance can be represented through a saliency map that highlights the contribution of each pixel at the prediction.

LIME creates a local, linear, surrogate model around the example for which an explanation is being created. The linear model is created by feeding perturbed versions of the example to the black box model and observing the outputs. For images, the perturbation of the input is achieved by dividing the examples into super-pixel regions, which are groups of neighboring pixels with similar color and brightness. Then a collection of synthetic images is created by replacing random super-pixels of the original image with a uniform, possibly neutral, color. The coefficients of the linear model reflect the importance of each feature to the ultimate prediction.

SHAP decomposes the prediction for a given example into the contribution from each individual feature. To estimate the contribution of each feature, perturbed versions of the example are fed to the black box model, and the corresponding outputs are observed. This decomposition is not based on linearizing, and so the parameter rankings are consistent with the original model.

In this section we also give an overview of GRADCAM (Gradient-weighted Class Activation Map) that is a post-hoc local explainer for image data [18] that differently from LIME and SHAP is model specific. As it will be explained in Section 3.1 to address the challenge of aligning semantics of the machine generated explanation with human interpretation, we describe the creation of a surrogate white-box approach as a stand-in. Towards that end we built a simple prototype that uses the inputs and outputs of a black box model to train a second explainer model. This second model would serve as a surrogate for the black box model, but since it is *not* black box other algorithms can be used to produce explanations. GRADCAM, based on gradient information of a Convolutional Neural Network (CNN), assigns each neuron a saliency value for the decision of interest and backpropagates this information to the last convolutional layer of the CNN.

Approaches like LIME and SHAP generate low-level explanations that in case of imagery inputs are expressed as the presence/absence of pixels that were most important to make a particular decision. However, for AI-aided decision-making systems, the human-in-the-loop that must evaluate the AI decision, usually a domain-expert without technical knowledge in AI, prefers high-level and concept-based explanations that can easily understand and reason with. Such high-level and concept-based explanations are more familiar to humans and more aligned with the human' internal representation of the decision-making problem. Section 2.2 will provide an overview of relevant methods that generate concept-based explanations.

2.2 Methods based on Concept Attribution

In this section we review some of the most relevant methods that generate explanations based on concept attribution. Here concept attribution refers to the ability of a method to quantify how much a concept has contributed to a particular class prediction. Concept-based explanations have the stated goal of using concepts that are "meaningful", "coherent", and have "importance" [20]. Here meaningful is defined as having intrinsic meaning to a human; coherent indicates that examples of concepts should be similar to each other and distinct from other concepts; importance signifies that the presence of a concept is necessary to true prediction from examples in a certain class.

TCAV (Testing with Concept Activation Vector) [19] quantifies the degree to which hand-selected concepts are important to a classification result. Every concept is represented by a Concept Activation Vector (CAV) that is built by interpreting post-hoc a neural network's internal state in terms of such hand-selected concepts. However, approaches like TCAV require a human to provide hand-labeled

examples of concepts. This might introduce human bias in the explanation process by failing to choose the right concepts. Moreover, methods like TCAV that can only leverage concepts that are already labeled and identified by humans, have limited power in discovering other relevant attributes [20].

ACE (Automated Concept-based Explanation) [20] is an evolution of TCAV that instead of human-labeled concept data, can automatically discover concepts by breaking up images into segments (e.g., super-pixels) and by seeing which ones are clustered in the representation space. Then, like in TCAV, ACE quantifies how much these clusters contributed to the prediction of a class.

Similarly to ACE, ConceptSHAP [21] targets having concepts consistently clustered to certain coherent spatial regions. For each discovered concept an importance score is defined from a set of Concept vectors (Cs) by utilizing Shapley values.

StylEx [22] introduces a method for automatically discovering visual interpretable attributes and use them for counterfactual explanations. Counterfactual explanations provide alternative inputs, where a small set of attributes is altered, and the different classification outcomes are observed. For each discovered attribute, StylEx generates a counterfactual example showing how manipulating such attribute affects the classifier prediction (*Had the input x been $\neg x$ then the classifier output would have been $\neg y$ instead of y*). To generate visual counterfactual explanations StylEx leverages a StyleGAN [23] trained to discover classifier-related attributes.

The concept-based explanation literature presented in this section is lacking in some respects. On one hand, techniques like TCAV require resources to generate human hand-labeled examples of concepts and its discovery power is rather limited. On the other hand, techniques like ACE and StylEx can automatically discover relevant concepts but there is no guarantee that such concepts will be human interpretable. Research related to such methods have only found evidence that the discovered concepts *tend* to be meaningful to people.

Relative to the literature described in Sections 2.1 and 2.2, there are still many research challenges that need to be addressed in pursuit of creating user-intuitive and effective explanations of black box machine learning models. Section 3 discusses key areas requiring further investigation for improving the quality of model-agnostic explanations.

3 Toward User-Intuitive Explanations

Treating the model as a black box while producing explanations is convenient because it allows the predictive and explanation systems to be entirely decoupled. However, most black box explanation methods are currently reliant on the inherent interpretability of the features that they use. That is, they assume that the semantics of the explanations are aligned with human interpretations. The ease of interpreting the explanation is directly related to the extent to which the features are interpretable. This may work generally, but in performance-critical applications a more intentional selection of semantically meaningful features is desirable, perhaps necessary. Such an approach would help ensure that produced explanations would be more meaningful to an operator. To address this challenge of aligning semantics of the machine generated explanation with human interpretation, we first describe the creation of a surrogate white-box approach as a stand-in (Section 3.1), and subsequently describe the concept of a semantic alignment method (Section 3.2) based on a neuro-symbolic approach.

3.1 Creating a Surrogate White Box Model

Methods like LIME and SHAP reviewed in Section 2.1 are limited to ranking feature importance (although "features" can be quite a flexible concept). One way to address this is to use the inputs and outputs of the black box model to train a second explainer model. This second model would serve as a surrogate for the black box model, but since it is *not* black box other algorithms can be used to produce explanations.

A simple prototype of this system was created to investigate the possibilities. The main goal was to create a white-box model based on outputs from the black box model. This was done in the following way: first, we used labeled training data from the Rareplanes dataset; specifically, images of aircraft along with their classification of how many engines they have (an integer 0 through 4). This set of images was then used to train a convolutional neural network (CNN) to use those cropped images and classify the number of engines. Finally, a white-box, model-aware, explanation method known as Grad-CAM [18] was used to produce saliency maps illustrating the regions of images that were most influential for making a classification. This process is depicted in Figure 2.

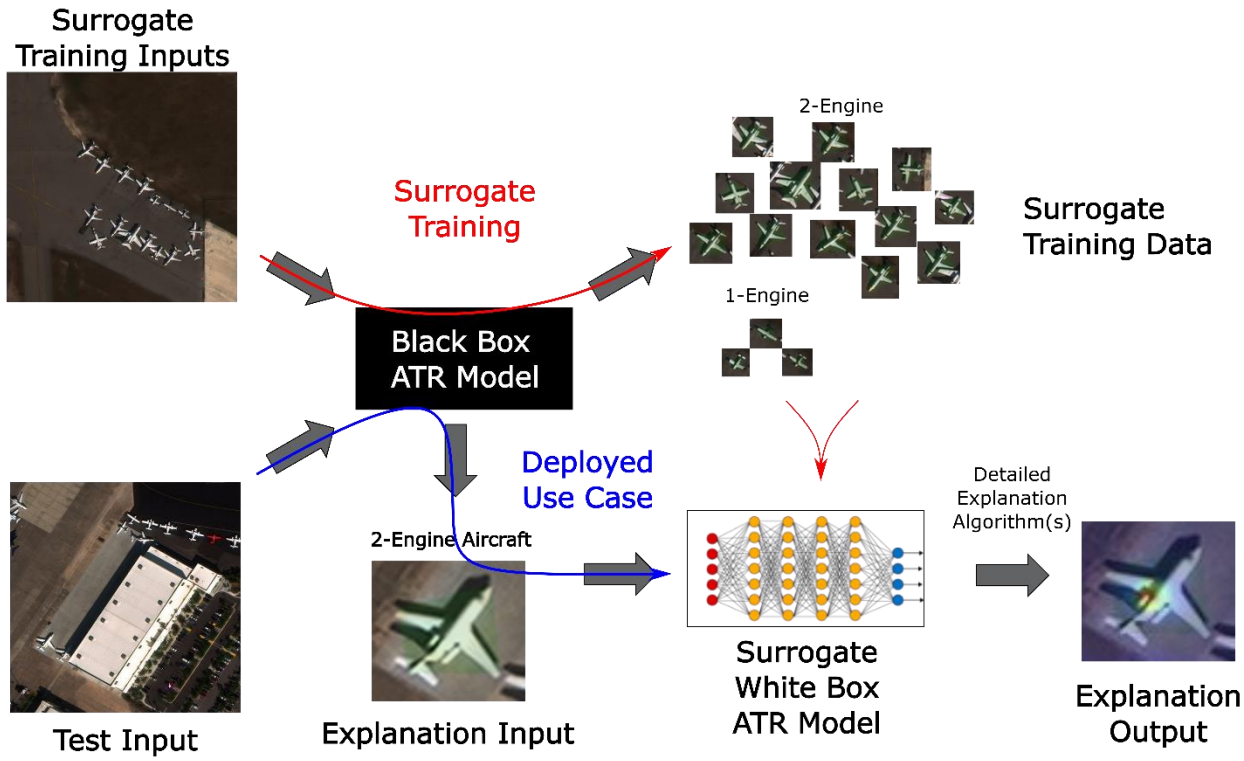

Figure 2 -- Diagram illustrating the training process for the surrogate model as well as an example of how the model could be used after being deployed.

The surrogate white box model used for this prototype was based on the VGG-16 model originally designed for the ImageNet dataset [24]. The architecture of the original VGG-16 model is shown in Figure 3; for our prototype specialized model a modified version of the VGG-16 model was created (Figure 4). In essence convolutional layers 1-4 were "frozen" (kept their weights from being trained on ImageNet). The 5th convolutional layer had its weights re-trained on the surrogate training data set discussed in Figure 2. The final fully connected layers had their sizes changed to better fit the specific classification problem. Note that this description highlights what we did for the prototype system that we built and is not meant to be generally prescriptive for future work; it simply demonstrates that it is possible to build a surrogate white box model as a means to utilize different non-black box explanation generation algorithms.

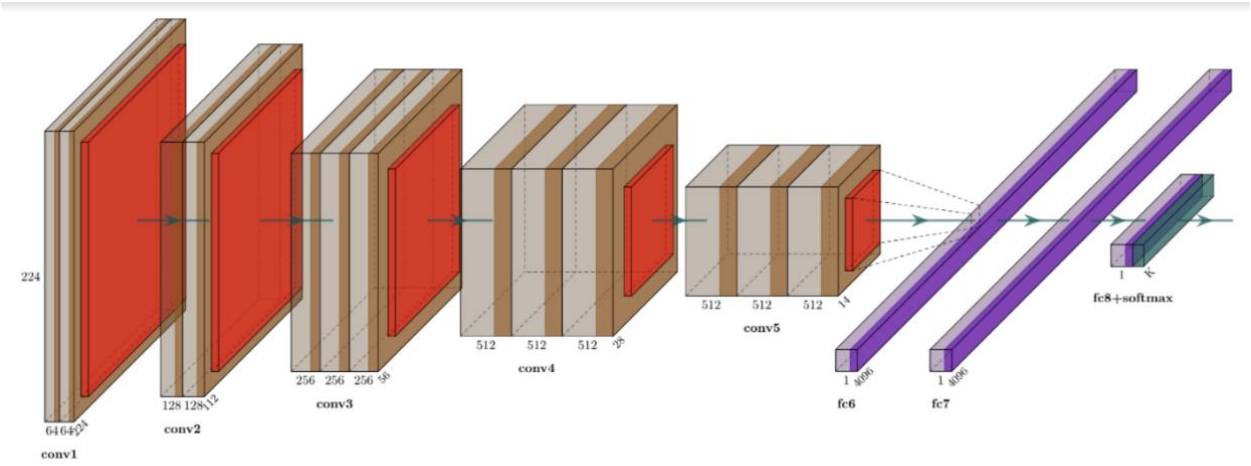

Figure 3 -- Diagram showing the layout of the VGG-16 architecture [24].

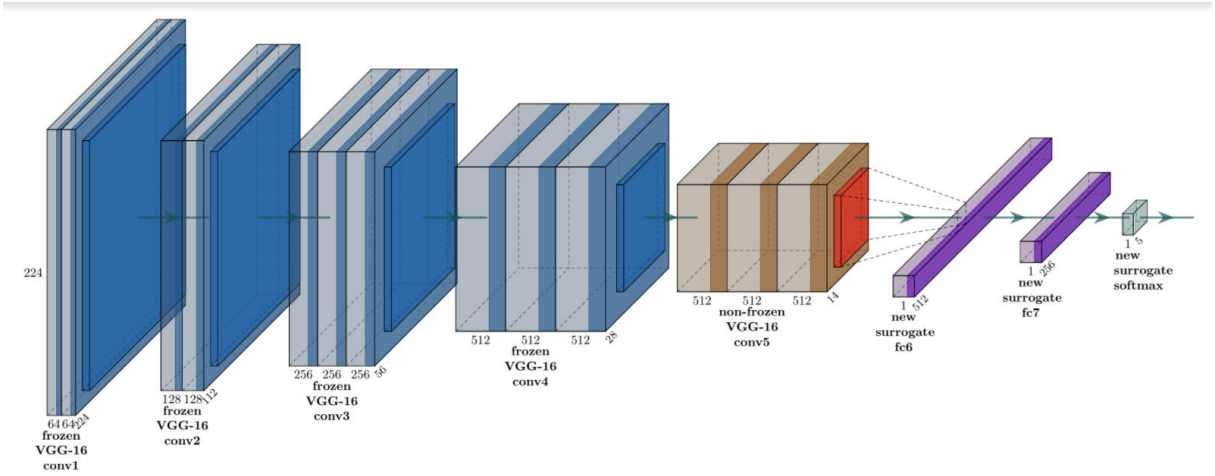

Figure 4 -- Surrogate white box model reuses the architecture from the original VGG-16 model. Layers 'conv1', 'conv2', 'conv3', and 'conv4', were left unmodified (i.e., "frozen"). The 'conv5' layer kept the same structure, but the weights were modified after training on the labeled images discussed. Layers, 'fc6' and 'fc7' had their size changed; 'fc8' was removed, and the final 'softmax' layer was changed to accommodate the 0-4 engine classification output of the Rareplanes dataset.

We then applied the Grad-CAM explanation method to produce saliency maps illustrating the regions of images that were most influential for making a classification. It is important to note that while we used Grad-CAM in this prototype, it is also possible to apply a host of other explanation methods. This serves as a proof of concept that it's possible to train a surrogate white box model that enables the application of non black box explanation algorithms to explain the decisions of a black box model. Figure 5, and Figure 6 contain example explanations of a surrogate white-box model trained to mimic the Rareplanes classifications of the number of engines as illustrated in Figure 2. Examples in Figure 5 generally demonstrate promising behavior, those in Figure 6 highlight some of the existing limitations. Figure 7, and Figure 8 contain results from similar surrogate models created to explain the role of the aircraft, and to classify the number of tail fins respectively.

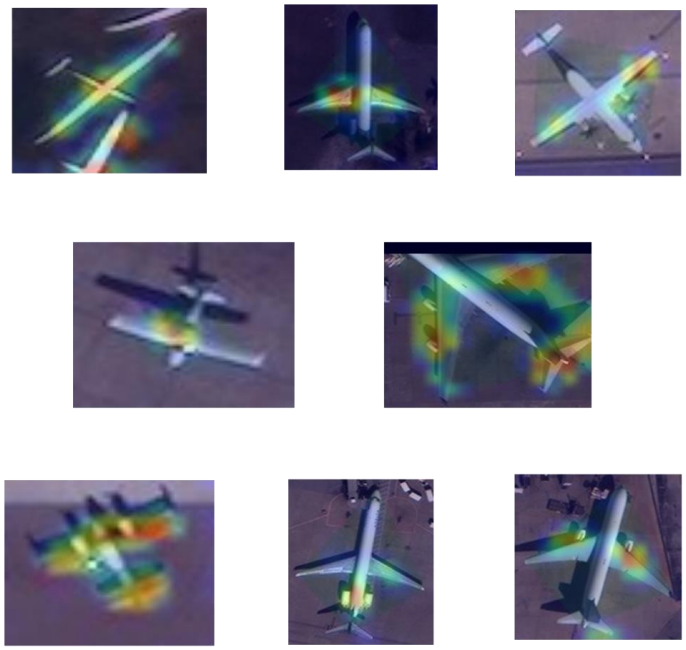

Figure 5 -- Selected Grad-CAM results from the surrogate white box model. The images uniformly indicate that the important areas for classifying the number of engines are the wings, tail, and engines of a plane.

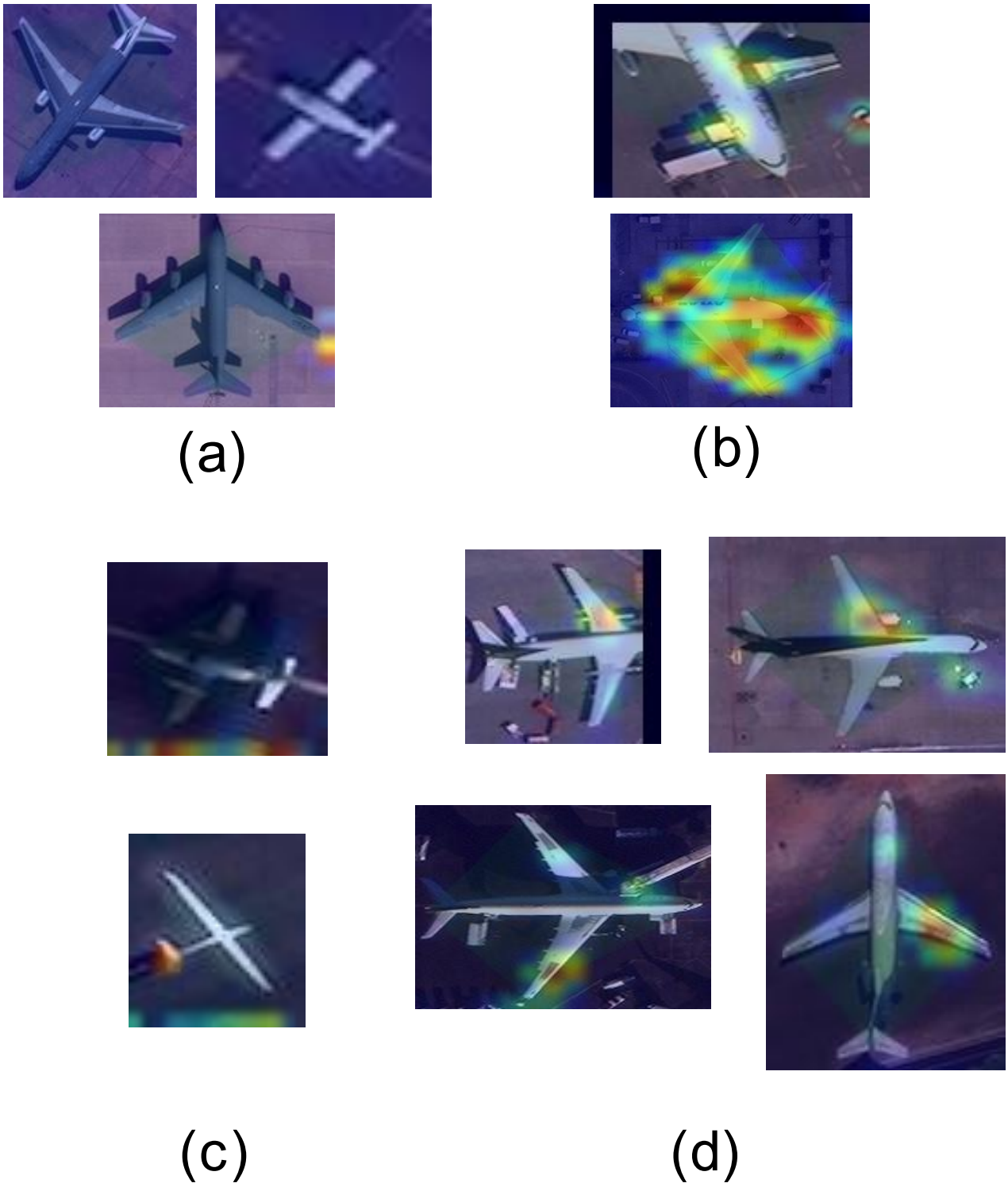

Figure 6 -- Selected results from Grad-CAM analysis of the surrogate model. These images represent unexpected results sometimes given by the algorithm. In (a) we see that sometimes the gradient is not very informative. The images in (b) highlight that sometimes the algorithm focused on too many features, or seemingly irrelevant features. Sometimes the gradient was at the edge of an image with almost not overlap on the plane (c). Finally, at other times the model seems to focus primarily on a single wing (d).

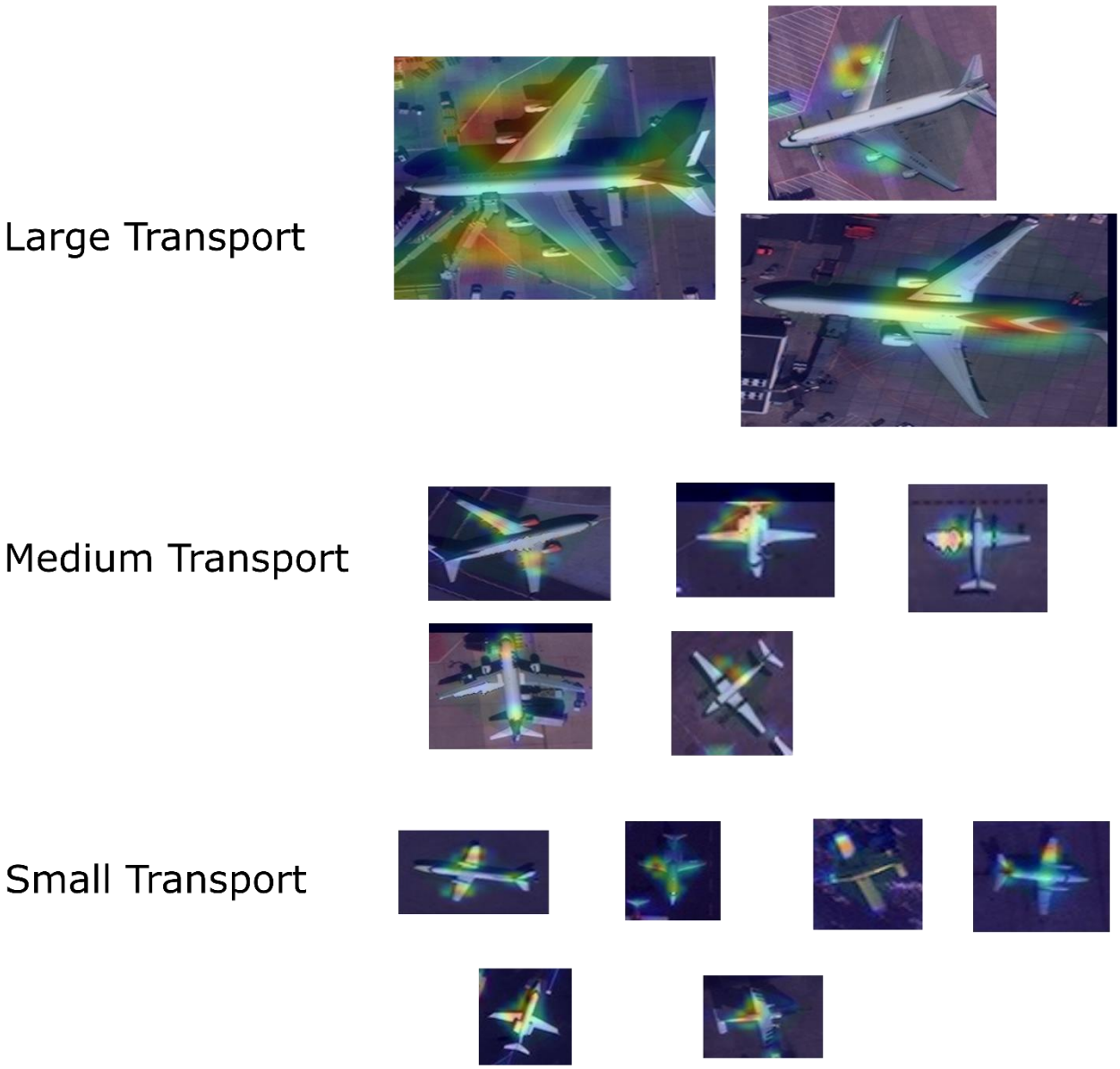

Figure 7 – Example explanations for prediction of civil role. The Rareplanes dataset has very few military aircraft (and thus has poor performance) so we don't display those results.

Figure 8 – Example explanations for a surrogate white-box classifier that predicts the number of tail fins. This surrogate model doesn't seem to be working correctly. Notice that there are few images that have hot-spots where GradCAM shows a high gradient.

3.2 Semantically Aligning the Explanation Model: A Neuro-symbolic Approach

Current machine learning methods can provide high classification accuracy but low human interpretability because the intermediary "hidden" layers are abstract statistical transformations. Indeed, in ML-aided decision-making tasks, the human-in-the-loop, usually a domain-expert without technical ML knowledge, prefers high-level concept-based explanations instead of low-level explanations based on model features. In order to create meaningful explanations for a black box model, that can align with user' internal representation of the problem, concept-based explanations can be leveraged. However, as discussed in Section 2.2 current methods based on concept attribution either require resources to generate hand-labeled examples of concepts or, in case such attributes can be automatically discovered, there is no guarantee that such concepts will be human interpretable. Most recently neuro-symbolic AI has been proposed as tool for achieving more explainable, transparent, and trustworthy systems. The neuro-symbolic approach is mostly motivated by the complementary strengths of sub-symbolic computations (i.e., data-driven AI) and symbolic representations (i.e., knowledge-driven AI) that could lead to designing hybrid, more intelligent AI systems [25][24]. More specifically, the neuro-symbolic approach offers more explainable, transparent and trustworthy systems by complementing sub-symbolic approaches and their ability to deal with large amounts of data, to handle noise, and to capture the richness of perceptual data, with symbolic methods, allowing to encode knowledge in the form of language-like, structured propositions that can be endlessly recombined to allow high-level reasoning across tasks and domains [26].

Knowledge graphs have been proposed as tools to support the explanation of machine learning models [26],[27],[28]. In the context of image classification tasks, early work focused on manually creating an ontology capturing spatial concepts, colors, textures, and their relationships, and incorporating it in an object recognition classifier [29]. The domain knowledge captured in the ontology was leveraged to foster model transparency acting as a user-friendly intermediate between the classifier and the end-user [29]. Most recently, existing large-scale and open-source knowledge graphs, such as OpenCyc[1], ConceptNet[2] and DBpedia[3], have been embedded in machine learning models to increase model transparency [30],[31],[32]. For example, background knowledge from ConceptNet was leveraged to explain objects in images with associated sentences in the form of captions [33]. A sentence-based image retrieval problem was leveraged to demonstrate that keywords in the captions that did not have a visual detector available, could be explained by leveraging the concepts and relations in the knowledge graph connected to such keywords.

Most of the neuro-symbolic approaches proposed so far do require a manual step of extracting knowledge sources. Indeed, one open challenge in knowledge-based explainable systems is the ability to automatically extract knowledge from graphs. In current approaches the identification of the correct portion of information in the graph to generate explanations requires using a human expert.

4 Metrics for Explanation Evaluation

This section describes relevant metrics that can be leveraged for measuring the effectiveness of the explanations generated by the techniques described in the previous sections of this report. For our purposes, the effectiveness of explanations must be assessed by evaluating how helpful the explanation is to the human end user. Such assessment requires human-in-the-loop experiments based on a variety of metrics and fall under the category of Human-grounded Metrics [34].

The DARPA XAI program formulated a model of explanation inspired by psychological principles [35]. The model describes the process underlying the interaction between a human end user and an explainable AI system as an opportunity to introduce metrics for assessing the human + AI system performance. Following the DARPA XAI model, initial instructions on how to use an AI system enable a human end-user to form an early mental model of the task and the AI system. As the end-user starts using the AI system, explanations are generated to help the human refine his/her mental model. This in turn should take to better performance of the human + AI system, and to the establishment of appropriate trust and reliance in the AI system. To evaluate the performance of the human + AI system, five measurement categories were proposed in [35].

The "*Goodness Criteria*" provides a means to assess the goodness of explanations, based on factors revolving around clarity and precision. A "goodness checklist" that can be used to either design "goodness" into the explanations, or to evaluate the a-priori goodness of the explanations that the system generates is described.

The "*Test of Satisfaction*" enables the assessment of whether users are satisfied by the received explanations. This is an a-posteriori judgment of explanations that leverages terms such as understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness, included into a Likert scale for review and evaluation.

[1] http://www.cyc .com /opencyc /a
[2] http://conceptnet .io/
[3] https://www.dbpedia .org

The "*Test of Understanding"* gives information on how well users understand the AI systems (i.e., mental model). Methods to elicit human users' mental models, resulting in data that can be easily scored, categorized, or analyzed are described. The intent of such methods is to provide some sort of structure or "scaffolding" that supports the user in explaining their reasoning process while solving a task. Examples of eliciting users' mental models include a "*Retrospection Task*" for which probe questions are presented to participants about their reasoning just after the reasoning task has been performed (e.g., "*Can you describe the major components or steps in the [software system, algorithm]?*"). Another example of eliciting users' mental models is a "*Prediction Task*" during which users are presented with test-cases, and they are asked to predict the results and then explain why they thought they would obtain those results (e.g., "*What will the [software system, algorithm] do next?", "How do I intervene?").* In a *"Self-explanation Task"* users express their own understanding by generating self-explanations. This helps learners to refine their knowledge. Last, a *"Think-aloud Solving Task"* participants think aloud while they solve a task.

The *"Test of Performance* provides information on how the human-AI system perform with the intent to understand if there is an improvement in the user's decision and task performance. According to this model, the user performance will improve as a result of receiving satisfying explanations, and it will be a function of the quality of his/her mental model.

The last measurement proposed in the DARPA XAI program is about "*Appropriate Trust and Reliance"* providing information on whether the user's trust and reliance on the AI system are appropriate. Currently, the assessment of trust in such systems is predominantly done via human self-reported behavior. However, rather than looking at scales for measuring interpersonal trust, the interest here was in scales designed to assess human trust in AI systems. Typical scales include questions to assess trust (e.g., "*Do you trust the machine's outputs?*") and reliance (e.g., "*Would you follow the machine advice?"*). Other available scales are highly specific to the context and application of interest (e.g., assistive robotic technology for elderly). A trust scale that incorporates items from other relevant scales and that might be used in the XAI context has been defined in [35]. According to the outcome of the XAI Program, trust assessment should be a repeat measure that requires multiple measures taken over time and integrated for overall evaluations of human user + system performance.

4.1 Future Research Needs

For the research described in Section 3, new effective evaluation metrics based on trustworthiness and acceptance will need to be defined. Future research directions in this area can incorporate technologies for continuous monitoring and automatic assessment of human trust in AI systems. Indeed, trust depends on time-varying factors that can influence the human decision-making process during interactions with AI systems [36]; the intended user should learn under what conditions the system fails and it succeeds in accomplishing the user's goals [7]. Further, trust may also be built or calibrated "on the job" as operation time increases, facilitated by the incorporated continuous monitoring and automatic assessment technologies. Interesting directions to explore include leveraging psychophysiological measurements to collect data that, when used as input to classification algorithms, map continuous data to a categorical trust level.

5 Conclusion

This work has focused on methods for making AI/ML models, leveraged by Automatic Target Recognition systems for example, more interpretable to human operators by employing AI explainer generation technologies. In ATR it is necessary for the human operator to be able to verify that

identified targets meet the necessary criteria; user-intuitive and accurate explanations will help them to do so. Model-agnostic and post-hoc algorithms were identified as promising technologies. There are still several technical challenges that exist before this technology can be considered capable of meeting all desired characteristics. Approaches like LIME and SHAP generate low-level explanations that in case of imagery inputs are expressed as the presence/absence of pixels that were most important to make a particular decision. However, for AI-aided decision-making systems, the human-in-the-loop that must evaluate the AI decision, usually a domain-expert without technical knowledge in AI, prefers high-level and concept-based explanations that can easily understand and reason with. Such high-level and concept-based explanations are more familiar to humans and more aligned with the human' internal representation of the decision-making problem. The concept-based literature has made significant progress in methods that can automatically discover relevant concepts and quantify how much such concepts contribute to a particular class prediction. However, there are still challenges that need to be addressed. On one hand, techniques like TCAV require resources to generate human hand-labeled examples of concepts and its discovery power is rather limited. Techniques like ACE and StylEx can automatically discover relevant concepts but there is no guarantee that such concepts will be human interpretable. Research related to such methods have only found evidence that the discovered concepts *tend* to be meaningful to people. To address this challenge of aligning semantics of the machine generated explanation with human interpretation, we describe the concept of a semantic alignment method and approaches that can be leveraged for that. Recently neuro-symbolic AI has been proposed as tool for achieving more explainable, transparent, and trustworthy systems. The neuro-symbolic approach is mostly motivated by the complementary strengths and weaknesses of sub-symbolic computations (i.e., data-driven AI) and symbolic representations (i.e., knowledge-driven AI) that could lead to designing hybrid, more intelligent AI systems. We also describe relevant metrics that can be leveraged for measuring the effectiveness of the explanations generated by the techniques described in this work. For our purposes, the effectiveness of explanations must be assessed by evaluating how helpful the explanation is to the human end user. Such assessment requires human-in-the-loop experiments based on a variety of metrics and fall under the category of Human-grounded Metrics. For the research described in this work, new effective evaluation metrics based on trustworthiness and acceptance will need to be defined. Future research directions in this area can incorporate technologies for continuous monitoring and automatic assessment of human trust in AI systems.

6 Acknowledgments

This project was funded by Raytheon Intelligence & Space Internal Research and Development (POCs Frank Tanner and Laura Strater).

7 References (APA formatting)

[1] Bocskor, A., Hunyadi, M., & Vince, D. (2017). National Academies of Sciences, Engineering, and Medicine (2015) The Integration of Immigrants into American Society. Washington, DC: The National Academies Press. 458 pages. INTERSECTIONS: EAST EUROPEAN JOURNAL OF SOCIETY AND POLITICS, 3(3), 157-161.

[2] Endsley, M. R., Bolté, B., & Jones, D. G. (2003). Designing for situation awareness: An approach to user-centered design. CRC press.

[3] Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). Situation awareness-based agent transparency. Army research lab aberdeen proving ground md human research and engineering directorate.

[4] Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters, 2(4), e61.

[5] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

[6] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[7] Board, D. I. (2019). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense: Supporting Document. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

[8] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18.

[9] Rareplanes: Synthetic data takes flight. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 207-217).

[10] Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245-317.

[11] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.

[12] Mittelstadt, B., Russell, C., & Wachter, S. (2019, January). Explaining explanations in AI. In Proceedings of the conference on fairness, accountability, and transparency (pp. 279-288).

[13] Joshi, G., Walambe, R., & Kotecha, K. (2021). A review on explainability in multimodal deep neural nets. IEEE Access.

[14] Vermeire, T., Laugel, T., Renard, X., Martens, D., & Detyniecki, M. (2021). How to choose an Explainability Method? Towards a Methodical Implementation of XAI in Practice. arXiv preprint arXiv:2107.04427.

[15] Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2021). Benchmarking and survey of explanation methods for black box models. arXiv preprint arXiv:2102.13076.

[16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[18] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-CAM: Why did you say that? arXiv preprint arXiv:1611.07450.

[19] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018, July). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In International conference on machine learning (pp. 2668-2677). PMLR.

[20] Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. Advances in Neural Information Processing Systems, 32.

[21] Yeh, C. K., Kim, B., Arik, S., Li, C. L., Pfister, T., & Ravikumar, P. (2020). On completeness-aware concept-based explanations in deep neural networks. Advances in Neural Information Processing Systems, 33, 20554-20565.

[22] Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., ... & Mosseri, I. (2021). Explaining in Style: Training a GAN to explain a classifier in StyleSpace. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 693-702).

[23] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8110-8119).

[24] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[25] Van Harmelen, F., & Teije, A. T. (2019). A boxology of design patterns for hybrid learning and reasoning systems. arXiv preprint arXiv:1905.12389.

[26] Tiddi, I., & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. Artificial Intelligence, 302, 103627.

[27] Lecue, F. (2020). On the role of knowledge graphs in explainable AI. Semantic Web, 11(1), 41-51.

[28] Futia, G., & Vetrò, A. (2020). On the integration of knowledge graphs into deep learning models for a more comprehensible AI—Three Challenges for future research. Information, 11(2), 122.

[29] Maillot, N. E., & Thonnat, M. (2008). Ontology based complex object recognition. Image and Vision Computing, 26(1), 102-113.

[30] Sarker, M. K., Xie, N., Doran, D., Raymer, M., & Hitzler, P. (2017). Explaining trained neural networks with semantic web technologies: First steps. arXiv preprint arXiv:1710.04324.

[31] Daniels, Z. A., Frank, L. D., Menart, C. J., Raymer, M., & Hitzler, P. (2020, April). A framework for explainable deep neural models using external knowledge graphs. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II (Vol. 11413, p. 114131C). International Society for Optics and Photonics.

[32] Marino, K., Salakhutdinov, R., & Gupta, A. (2016). The more you know: Using knowledge graphs for image classification. arXiv preprint arXiv:1612.04844.

[33] Icarte, R. T., Baier, J. A., Ruz, C., & Soto, A. (2017). How a general-purpose commonsense ontology can improve performance of learning-based image retrieval. arXiv preprint arXiv:1705.08844.

[34] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[35] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

[36] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

[37] Brett W. Israelsen and Nisar R. Ahmed. 2019. ``Dave...I Can Assure You...That It's Going to Be All Right..." A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships. ACM Comput. Surv. 51, 6 (January 2019), 113:1–113:37.